Kulcsszókinyerés alapú dokumentumklaszterezés

Berend Gábor[1], Farkas Richárd[1], Vincze Veronika[2],
Zsibrita János[1], Jelasity Márk[2]

[1]Szegedi Tudományegyetem, TTIK, Informatikai Tanszékcsoport
Szeged, Árpád tér 2., e-mail:{berendg, rfarkas, zsibrita}@inf.u-szeged.hu
[2]Magyar Tudományos Akadémia, Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103., e-mail:{vinczev, jelasity}@inf.u-szeged.hu

Kivonat A szöveges dokumentumok lényegi mondanivalóját tömören összegezni képes kifejezések kitüntetett fontossággal bírnak: számos nyelvtechnológiai alkalmazás profitálhat ismeretükből a katalogizáló és kivonatoló rendszerekben történő felhasználásuktól kezdve egészen az információ-visszakereső alkalmazásokig. Cikkünkben automatikusan meghatározott kulcsszavak minőségét alternatív módon, egy dokumentumklaszterező alkalmazásban való felhasználásuk kapcsán vizsgáltuk. A munkánk során felhasznált dokumentumokat a *Magyar Számítógépes Nyelvészeti Konferencia (MSzNy)* megjelent konferenciakötetеinek cikkei képezték. A cikkekből történő csoportképzést összehasonlítottuk a cikkekben előforduló n-gramok, valamint gépi tanulás útján meghatározott kulcsszavak alapján is. Eredményeink tükrében kijelenthető, hogy a kulcsszavak hasznosak a dokumentumklaszterezés feladatának megsegítésében is. A cikkek automatikus kulcsszavai alapján értelmezett hasonlósági gráf vizualizálása és klaszterezése során tapasztaltak alapján megfigyelhető volt továbbá a nyelvtechnológia egyes részterületeinek elkülönülése, időbeli fontosságuk változása, amely alapján az automatikus kulcsszavak – alkalmazásoldali szempontból – megfelelő minőségére következtethetünk.

Kulcsszavak: automatikus kulcsszókinyerés, dokumentumklaszterezés

1. Bevezetés

A dokumentumokhoz – automatikusan avagy manuálisan – rendelt kulcsszavak azon túl, hogy egy tömör összefoglalójaként értelmezhetők az egyes dokumentumoknak – és ezáltal alkalmassá válnak azok visszakeresésének vagy osztályozásának megkönnyítésére –, fontos eszközei lehetnek a dokumentumok közötti hasonlóságok meghatározásának. Jelen cikkben azt a kérdést vizsgáljuk, hogy a dokumentumok között definiált hasonlósági reláció modellezésére alkalmasabb-e az egyes dokumentumok kulcsszavaira támaszkodni, mint a hagyományos vektortérmodellre (ahol a dokumentumokat a bennük előforduló összes n-grammal jellemezzük).

2. Kapcsolódó munkák

Az elmúlt években számos tudományos eredmény látott napvilágot hazai és nemzetközi szinten egyaránt a dokumentumok lényegét leírni hivatott kifejezéshalmazok automatikus meghatározását végző rendszerekre nézve. Ezen munkák jellemzően angol nyelvű tudományos publikációk kulcsszavainak automatikus meghatározását tűzték ki célul (pl. [1],[2] és [3]), azonban akadnak kivételek is, amelyek más doménű dokumentumok kulcsszavazására vállalkoztak (pl. [4], [5] és [6]). Mindamellett, hogy az angol nyelvű tudományos publikációkból történő kulcsszókinyerésnek tehát igen bő irodalma áll rendelkezésre, magyar nyelvű politika- és neveléstudományi témában íródott tudományos publikációk kulcsszavainak gépi tanuláson alapuló meghatározására is született már kísérlet [7].

A korábbi munkák hatékonyságának objektív megítélésének komoly gátat szab az a tény, hogy a kulcsszavak minőségének emberi elbírálása meglehetősen szubjektív, valamint az automatikus (szigorú sztringegyezésen alapuló) kiértékelésük szintén nehézségekbe ütközik az azonos (szinonim) vagy közel azonos (hipo- vagy hipernim) jelentésű kifejezések megjelenési formáinak sokszínűsége kapcsán. Jelen munka egyik célja egy alternatív kiértékelési lehetőség definiálása a kulcsszavak minőségének megítélésére, amely során a kulcsszavazás hatékonysága azon keresztül kerül lemérésre, hogy milyen mértékben sikerül egy korpuszt alkotó dokumentumokat elkülöníteni egymástól, csupán a hozzájuk tartozó legmegfelelőbbnek ítélt kulcsszavak ismeretének fényében.

A tudományos trendek természetesnyelv-feldolgozási eszközökkel történő kutatásának témájában szintén születtek már korábbi munkák. Ezek közül egy [8], ahol kulcsszavakhoz hasonló kifejezések előfordulásainak időbeli változását nyomon követve határozták meg a különböző tudományos résztémakörök relatív fontosságának változását.

3. Módszertan

A következő alfejezetek azt mutatják be, hogy az MSzNy-cikkarchívum egészének automatikus kulcsszavazása miként zajlott, majd ezt követően az egyes cikkekhez rendelt kulcsszavak alapján hogyan lettek azonosítva az egyes számítógépes nyelvészeti részterületek.

3.1. Automatikus kulcsszavazás

Mivel a cikkek szerzői csupán az esetek elenyésző hányadában látják el írásukat az azt jellemző kulcsszavakkal, ezért ahhoz, hogy a dokumentumok klaszterezése az őket legjobban leíró kulcskifejezések alapján is megtörténhessen, szükség volt egy olyan modell építésére, amely képes a kulcsszavak cikkek szövegéből történő automatikus kinyerésére. A feladat megoldása alapvetően a [7] által ismertetett módszert követte. A kulcsszavak meghatározására először a dokumentumból kigyűjtöttük a lehetséges kulcsszójelölteket, majd felügyelt tanulási módszerekkel azokat fontossági sorrendbe rendeztük. Jelen esetben a rangsorolás egy bináris

valószínűségi osztályozó a posteriori valószínűségein alapul, ahol a bináris osztályozót arra tanítjuk, hogy egy kulcsszójelölt szerepelt-e a dokumentum szerzője által a szóban forgó dokumentumhoz rendelt kulcsszavak között vagy sem. Ez a bináris tanuló a [7] jellemzőkészletéhez hasonló módon pozicionális, ortografikus és morfológiai jegyeik alapján reprezentálta a kulcsszójelölteket, az osztályozásukhoz pedig maximum entrópia modellt használtunk. A morfológiai elemzés elvégzésére a [9] modelljeit használtuk föl.

3.2. Dokumentumok hasonlóságának mértéke

Két dokumentum hasonlóságának mérésére több módszert is vizsgáltunk. Egyrészt ez a hasonlóság alapulhat az előző fejezetben bemutatott automatikus kulcsszavakon, vagy a dokumentum n-gramjain
($1 \leq n \leq 2$). Mindkét megközelítésre igaz, hogy egy dokumentumot a 10 „legjellemzőbb" kifejezésével írtunk le. Az n-gramok esetén a rangsoroló mérték a hagyományos tf-idf mutató volt, míg a kulcsszavakra támaszkodó reprezentáció esetében a bináris osztályozónk a posteriori valószínűsége volt mindez.

Két dokumentum esetén akkor beszélünk pozitív hasonlóságról, ha azok legalább egy közös „jellemző" kifejezéssel rendelkeznek. Két dokumentum kulcsszavaiból álló halmaz metszetének értékelésére több stratégiát is alkalmaztunk: egyes esetekben a mindkét halmazban megtalálható kifejezések fontosságértékének *maximumai*, *minimumai*, *átlagai*, *szorzatai*, illetve *harmonikus közepei* lettek véve, majd a két dokumentum globális hasonlóságának meghatározásához ezek az értékek összegezve lettek átfedésben álló kifejezéseik fölött. Két további megközelítés az átfedő kifejezések fontosságát nem, csupán azok számosságát vette figyelembe: ezek a *Dice*- és *Jaccard*-együtthatókon alapuló módszerek voltak.

3.3. Hasonlósági gráf alapú klaszterezés

Végső célunk egy dokumentumhalmaz klaszterezése, melyhez a csúcsaiban dokumentumokat reprezentáló (irányítatlan) hasonlósági gráfot építünk fel, a gráfban szereplő éleket pedig oly módon súlyoztuk, hogy azok értékei az előző fejezetben bemutatott páronkénti dokumentumhasonlóság-értékek voltak. Két dokumentumnak megfeleltethető a, b csúcs között csak akkor vezet él a gráfban, ha kulcsszavaik metszete nem üres, valamint az átfedés mértékét számszerűsítő súlyozás alapján b az a dokumentumhoz leghasonlóbb 3 dokumentum között szerepel vagy fordítva (a szerepel a b-vel legnagyobb hasonlóságot mutató 3 dokumentum között). A klaszterezést (particionálást) ezen a gráfon hajtjuk végre.

Egy adott gráfparticionálást jellemző modularitás [10] kiszámításával egy jósági értéket rendelhetünk a felbontás minőségére nézve, mely figyelembe veszi a gráf topológiájából adódóan az egyes csúcspárok között elvárható élek számát, valamint egy tényleges felbontás során az egyes csoportokon belül vezető élek tapasztalt számát:

$$Q = \frac{1}{2m} \sum_{ij} \Big(A_{ij} - \frac{k_i k_j}{2m} \Big) \delta(C_i, C_j), \tag{1}$$

amelyben az összegzés minden *lehetséges* élre (minden i és j csúcsra) vonatkozik, és ahol az A_{ij} a particionálandó gráf szomszédsági mátrixának egy eleme, m a gráfban található élek száma, az összegzésben található hányados pedig az i és j csúcsok összeköttetésének – k_i és k_j fokszámértékekre támaszkodva számított – várható értéke, a δ függvény pedig az ún. Kronecker-delta, mely akkor veszi fel az 1 értéket, ha az i és a j csúcsok megegyező klaszterbe soroltak, egyébként 0.

Egy gráf olyan felbontásának meghatározása, amely erre a mutatóra tekint maximalizálandó célfüggvénye alapjául, erősen $\mathcal{NP}$-teljes [11]. Több közelítő eljárás látott már azonban napvilágot a probléma minél hatékonyabb, gyors megoldására, melyek között találunk szimulált hűtéstől kezdődően spektrálmódszereken át mohó megközelítéseket alkalmazókat is.

A spektrálmódszereken alapuló eljárások hátránya a megfelelő skálázódásuk hiánya, noha az alkalmazásukkal elért eredmények gyakorta felülmúlják a más megközelítésekkel kapottakat. A [12] által javasolt mohó optimalizáló stratégia kifejezetten nagy gráfokon is működőképesnek bizonyult, így az általuk javasolt eljárást valósítottuk meg a dokumentumhasonlósági gráf particionálására. Jóllehet a kísérleteink során megkonstruált gráfok csúcsainak számai mindössze százas nagyságrendben mozogtak, abból kifolyólag, hogy a későbbiekben nagyságrendekkel nagyobb dokumentumkollekciókon is használható legyen az algoritmusunk, ezért fontosnak éreztük a particionálást elvégző eljárásnak olyat választani, amely kedvező számítási bonyolultsággal rendelkezik.

A [12] szerzői által javasolt megközelítés egy alulról-felfele építkező klaszterező eljárás, mely kezdetén minden csúcsot egy külön klaszterbe sorol, majd a további lépések alkalmával a csúcsok meglátogatása során azokat a lokálisan legjobb modularitásnövekményt eredményező közösséghez sorolják (esetleg egyikhez sem). Egy i csúcs C közösségbe történő mozgatása során kettős hatás figyelhető meg: egyrészt növeli a globális modularitás értékét azon élei által, amelyek immáron a C közösségbeli szomszédjaival való összeköttetést biztosítják, másrészről viszont a modularitás bizonyos mértékű csökkenése is megfigyelhető lesz azon élei kapcsán, amelyek a korábbi közösségének tagjaival való összeköttetésért voltak felelősek. Egy i csúcs C közösségbe történő átmozgatásának hatása a következők szerint összegezhető:

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right], \quad (2)$$

ahol $\sum_{in}$ és $\sum_{tot}$ értékek rendre a C közösségen belül, illetve a C közösséget érintő élek súlyainak összege, k_i és $k_{i,in}$ pedig rendre az i csúcsot tartalmazó, illetve az i csúcsot a C közösséggel összekötő élek súlyainak összege, m pedig a particionálandó gráfban található élek összsúlya. Miután minden csúcs besorolást nyert az egyes közösségekbe, az algoritmus a kialakult közösségeket összevonva, és azokat egy csúcsként kezelve megismétli az előző eljárást. Az előzőekben ismertetett eljárás gyorsaságán túl egy további előnye, hogy a kialakuló közösségek száma a particionálandó gráf topológiája alapján kerül meghatározásra, a meg-

1. táblázat. Az MSzNy legnépszerűbb témáinak eloszlása 2003-2013 között.

	2003	2004	2005	2006	2007	2009	2010	2011	2013	Összesen	Arány
cikkek száma	59	46	52	49	32	45	46	40	42	411	
morfológia	6	6	9	2	3	3	4	7	8	48	11,68 %
beszédfelismerés	5	5	5	4	5	7	6	4	2	43	10,46 %
pszichológia	5	7	5	10	6	5	0	3	2	43	10,46 %
szemantika	3	3	3	6	3	4	7	7	6	42	10,22 %
lexikográfia	7	4	6	2	0	4	6	4	5	38	9,25 %
szintaxis	5	4	7	2	5	2	5	3	3	36	8,76 %
korpusz	4	4	5	3	3	3	3	4	7	35	8,52 %
információkinyerés	2	4	2	3	1	7	10	1	5	35	8,52%
fordítás	6	7	3	4	1	4	1	4	1	31	7,54%
ontológia	1	1	4	9	0	2	1	1	0	19	4,62%

határozni kívánt csoportok számát egyéb eljárásokkal (pl. k-közép klaszterezés) szemben nem tekinti előre ismertnek.

4. Az MSzNy korpusz

Jelen munkában az *MSzNy* eddig megjelent konferenciaköteteinek cikkeinek klasztereződését vizsgáljuk meg. Az MSzNy-cikkeknél lehetőség van a szerzőknek kulcsszavakat megadni a cikkükhöz, amely lehetőséggel mindössze 45 esetben éltek a szerzők. Az előző fejezetben bemutatott felügyelt tanulási modellt ezen a 45 cikken tanítottuk.

A konferenciasorozat 2003-ban indult, és 2008 és 2012 kivételével minden évben megrendezésre került, így összesen kilenc év alatt megjelent 411 darab cikk képezte vizsgálódásaink alapját.Ahhoz, hogy a korpuszban megjelenő fő témakörök felügyelet nélküli detektálásának eredménye számszerűsíthető legyen, elvégeztük a korpuszba tartozó cikkek egy referenciabesorolását. Az emberi erővel történő témabesorolás alkalmával minden cikkhez az arra leginkább jellemző témakategóriák lettek meghatározva, mint például *morfológia*, *lexikográfia* stb. Arra törekedtünk, hogy a témakategóriák a számítógépes nyelvészet különféle részterületeit reprezentálják, így azok cikkekhez történő hozzárendelése felfogható legyen a dokumentumok egy osztályozásának.

Az MSzNy-cikkek kézi osztályozása és tematizálása lehetővé teszi azt is, hogy megvizsgáljuk, milyen trendek uralkodtak az utóbbi években a magyarországi számítógépes nyelvészet területén. Az 1. táblázat a tíz leggyakoribb tématerülethez társítható cikkek időbeli mennyiségi eloszlását mutatja. A táblázatból kiolvasható, hogy az összesítésben tíz leggyakoribbnak mutatkozó téma az összes, humán annotáció segítségével detektált témakör hozzávetőlegesen 90%-át fedi le. A táblázatból kiderül továbbá az is, hogy a megjelent cikkek számának tekintetében a legnépszerűbb téma a morfológia volt, valamint az is, hogy szintén számos cikk született a beszédfelismerés, illetve a pszichológiai szövegfeldolgozás témaköreiben.

Érdekes azt is megfigyelni, hogy az évek során hogyan alakult a különféle témák eloszlása. A morfológia a konferenciasorozat kezdetekor, illetőleg az utóbb években tölt be különösen előkelő pozíciót. A beszédfelismerés 2009 környékén volt népszerű téma a konferencián, a fordítás elsődlegesen 2003-2004 környékén, azaz a kezdetekben foglalt el dobogós helyet, a szemantika és a korpusznyelvészet előretörése viszont az utóbbi néhány évben figyelhető meg. Az információkinyerés különösen a 2009-2010-es években virágzott, legalábbis az MSzNy-es mutatók alapján. Kiugróan jó évnek bizonyult a 2006-os a pszichológiai szövegfeldolgozás és az ontológia számára. A táblázatban már nem szereplő tématerületek közül kettőt említünk meg: a 2007-es év különösen sok beszédszintézissel foglalkozó cikket hozott, illetőleg 2010 óta az információ-visszakeresés is egyre népszerűbb, azonban e témák az összesített helyezésük alapján nem kerültek a legjobb tízbe.

Az előző megfigyeléseket természetesen árnyalja annak ismerete, hogy csupán 9 kiadványon alapulnak, továbbá, hogy az MSzNy-en az egyes témákban évenként megjelenő cikkek számára kis elemszámú mintaként tekinthetünk csupán, melyek érzékenyek lehetnek a témák relatív népszerűségén kívüli egyéb tényezőkre is, ami azt eredményezi, hogy a minták statisztikai mutatói könnyedén módosulni képesek. Egy ilyen, a trendek megfigyelését megzavarni képes jelenség lehet például egy adott témájú projekt lezárulta, és az ezzel kapcsolatos disszeminációs tevékenységek megjelenése a konferencián, mely önmagában túlreprezentálttá képes tenni időszakosan egyes területeket.

A gépi feldolgozhatóság és a kiértékelés szempontjából azonban nem bizonyult minden cikk egyformán használhatónak, így az MSzNy archívumában található 411 cikk közül nem mind került felhasználásra a továbbiakban. Egyes cikkek idegen nyelven álltak csupán rendelkezésünkre, esetleg a dokumentumból történő szöveg kinyerése nem volt lehetséges az általunk használt eszközökkel, avagy duplikátumokkal volt dolgunk. Az előző okok miatt a hasonlósági gráfot így mindösszesen 394 dokumentum alkotta.

A kézi címkézés során egy dokumentum több kategóriamegjelölést is kaphatott, amennyiben az több számítógépes nyelvészeti részterületet is érintett. Az emberi osztályozás során bevezetésre került 31 témamegjelölés közül némelyek mindössze egy-egy ízben, akkor is csupán egy másik témamegjelöléssel karöltve lett fölhasználva, így fontosságuk igencsak megkérdőjelezhető volt. Az ilyen kevéssé fajsúlyosnak mondható témával rendelkező cikkeket – valamint az összes többi olyat is, ahol egy dokumentum témája nem volt egyértelműen meghatározott az emberi jelölés által – nem vettük figyelembe a kiértékelés során, vagyis amikor az automatizált kategorizálás átfedését vizsgáltuk az emberi osztályzásával. Ezen döntés meghozatalának hátterében az a megfontolás állt, hogy az ilyen cikkek esetében még az emberi többlettudás sem volt elegendő az egyértelmű témabesorolás meghozatalához, az általunk javasolt eljárás pedig éppen ilyen egyértelmű besorolásokat tesz.

Az előzőekkel összefüggésben 46 darab automatikus kulcsszóval egyébként ellátott – és ezáltal a hasonlósági gráfban is szerepeltetett – dokumentum nem képezte részét a korpusz cikkeinek közösségkeresés által meghatározott automatikus témabesorolásának kiértékelésében. Az eredetileg bevezetett 31 témakörből

2. táblázat. Az automatikus témamegjelölés során felhasznált cikkek témáinak eloszlása.

Téma	Mennyiség	Arány
pszichológia	40	14,04%
beszédfelismerés	38	13,33%
morfológia	32	11,23%
szemantika	32	11,23%
információkinyerés	30	10,53%
fordítás	27	9,47%
lexikográfia	25	8,77%
szintaxis	24	8,42%
korpusz	20	7,02%
ontológia	17	5,96%

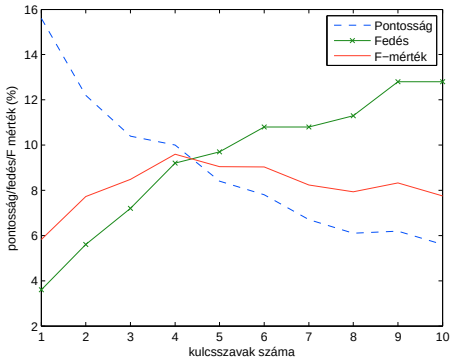
4 csupán más témák mellett kapott marginális szerepet, így a korpusz emberi kategorizálásra támaszkodó kiértékelésében is részt vevő dokumentumainak száma 337 volt, melyek 27 különböző egyedi kategóriába voltak sorolva. A 2. táblázatból kiolvasható, hogy a több kategóriába sorolt cikkek eltávolítását követően az egyes témamegjelölések hány alkalommal fordultak elő a kiértékeléshez használt adatbázisban. Megfigyelhető többek között az, hogy a megszűrt adatbázisban a leggyakoribb témának ezek után már a *pszichológia* mutatkozott, amit az okozott, hogy azon túl, hogy eredendően is viszonylag sok cikk lett hozzárendelve ehhez a kategóriához, ezek a témamegjelölések néhány kivételes esettől eltekintve teljesen egyértelműek is voltak, azaz esetükben az annotálás nem eredményezte további témák hozzárendelését a cikkekhez. Éles kontrasztot képez az előbbi témával a *morfológia* témaköre, amely előfordulásai harmadában valamely más témával együtt került megjelölésre.

5. Eredmények

Elsőként a kulcsszavazó modell hatékonyságát teszteltük, amikor is a 45 szerzői kulcsszóval ellátott dokumentum automatikusan kinyert kulcsszavainak minőségét ellenőriztük le 45-szörös keresztvalidációt alkalmazva. Egy kulcsszó elfogadása kizárólag abban az esetben történt meg, ha a normalizált alakra hozott kinyert kulcsszó tökéletes egyezést mutatott az adott cikkhez tartozó, és szintén normalizált alakban tárolt etalon szerzői kulcsszavak valamelyikével.

Megjegyzendő, hogy a 45 dokumentumhoz rendelt közel 200 kulcsszó közül mindössze 51,8% szerepelt ténylegesen is azokban a dokumentumokban, amelyekhez hozzá lettek rendelve, így a fedés értékének ez a lehető legmagasabb értéke az általunk használt kiértékelés mellett. Úgy gondoljuk azonban, hogy az eredmények ezen ténnyel való korrekciója után is a kapott számszerű eredményességi mutatók jóval elmaradnak attól a hasznosságtól, amellyel az automatikusan meghatározott kulcsszavak rendelkeznek. Mindezt arra alapozzuk, hogy a kifejezések egyezésének normalizált alakjaik szigorú sztringegyezésen alapuló vizsgálata sok szemantikai értelemben elfogadható kulcsszót álpozitív osztályba

sorolt: ilyenek voltak, amikor specializáló kifejezések nem kerültek elfogadásra a szigorú kiértékelés miatt (pl. a *felügyelt gépi tanulás* kifejezés a *gépi tanulás* ellenében), vagy amikor az elvárt és kinyert kulcsszavak jelentésükben egymással rokoníthatók voltak ugyan (adott esetben meg is egyeztek), ellenben írásmódjuk nem volt teljesen egyező (pl. a *morfológiai analízis* és *morfológiai elemzés* kifejezések).



1. ábra. A kulcsszavazó modell eredményessége a legvalószínűbbnek mondott $1 \leq k \leq 10$ kulcsszó tekintetében.

A továbbiakban már nem a kulcsszavak közvetlen minőségét, hanem használati értéküket vizsgáltuk egy dokumentumklaszterező felállásban, ahol a korpuszban szerepet kapó témákat kívántuk automatikusan meghatározni a dokumentumok szövege, illetve az abból kinyert kulcsszavak segítségével.

A cikkek által megkonstruált hasonlósági gráf particionálásának, valamint a cikkek ebből adódó automatikus témabesorolásának jóságát több mutatóval is jellemeztük. Egyrészt a közösségképzés végső minőségét számszerűsítő modularitási mutatóra támaszkodtunk. A dokumentumok particionálásának ezen mutatója csupán azt az aspektusát világítja meg az eljárásnak, hogy a hasonlósági gráfot mennyire sikerült az eredeti élstruktúrája függvényében megfelelő módon részgráfokra bontani. A megfelelőség foka azzal arányos, hogy az azonos közösségbe tartozó csúcsok között menő élek száma (vagy esetünkben azok súlyainak összege) minél nagyobb eltérést mutasson attól, mint amennyi él már csak a véletlennek is betudható lenne az egyes csúcsok fokszámai alapján.

A hasonlósági gráf magas modularitással történő felbontása azonban nem vonja feltétlenül maga után a meghatározott részkorpuszok szemantikus koherenciáját, ahogy ez a 3., valamint a 4. táblázatok kapcsán is észrevehető. Amennyiben ugyanis a csoportképződésért felelős élek olyan kulcsszavaknak köszönhetők, amelyek szemantikailag nem vagy csupán kevéssé köthetők egymáshoz, úgy kialakítható a gráf modularitás tekintetében kielégítő particionálása

3. táblázat. Automatikus kulcsszavakra nem támaszkodóan épített hasonlósági gráf particionálásának eredményei.

	Közösségek száma	Modularitás	Pontosság	V_1
Max	8	0,254	0,154	0,131
Min	9	0,372	0,160	0,127
Átlag	7	0,330	0,151	0,118
Szorzat	5	0,510	0,122	0,081
Harmonikus közép	9	0,336	0,175	0,142
Dice	2	0,071	0,113	0,025
Jaccard	2	0,072	0,113	0,025

olyan módon, hogy mindeközben a kialakult közösségek egymással rokonságba nem hozható elemekből állnak.

Éppen ezért szükségesnek éreztük további mutatók alkalmazását is a dokumentumok automatikus közösségekhez való társításának és az emberi erővel történő tematizálásuk átfedésének számszerűsítésére, ami érdekében több mutatót is alkalmaztunk. Az automatikus klasztereket leképeztük a kézzel jelölt különböző témaosztályokra, mely során mohó módon a még szóba jövő, legtöbb helyes besorolást eredményező klasztert rendeltük egy-egy etalon témaosztályhoz, amellyel egy injekciót határoztunk meg a közösségek és a témabesorolások között.

A kialakult csoportok szemantikus kohéziójának mérésére az információelméleti alapokon nyugvó V_1-mértékkel [13] is jellemeztük a kialakított közösségeket, amely a felügyelt tanulásból ismert F-mértékhez hasonlóan két érték harmonikus közepeként áll elő; a pontossággal és a fedéssel ellentétben itt a *homogenitás* és *teljesség* értékeket szokás definiálni. A homogenitás feltételes entrópiát használva számszerűsíti, hogy az egyes $c \in C$ közösségek mennyire diverzek a kézzel jelölt $k \in K$ témákhoz képest a

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{3}$$

képlet segítségével. A teljesség számítása analóg módon a

$$t = 1 - \frac{H(K|C)}{H(K)} \tag{4}$$

képlet alapján történik, a különbség mindössze annyi, hogy ennek esetében az etalon kategóriák diverzitása kerül számszerűsítésre a feltárt közösségek fényében. Egy tökéletes klaszterezés esetében az összes egy etalon témacsoportba tartozó elemet ugyanabban a megtalált klaszterben kell találjunk.

6. Diszkusszió

A 3. és 4. táblázatok összevetéséből kiderül, hogy minden tekintetben alkalmasabbnak bizonyult a hasonlósági gráf építése során csupán a dokumentumonkénti

4. táblázat. Automatikus kulcsszavak átfedése alapján épített hasonlósági gráf particionálásának eredményei.

	Közösségek száma	Modularitás	Pontosság	V_1
Max	12	0,689	0,303	0,365
Min	15	0,766	0,344	0,406
Átlag	14	0,763	0,300	0,391
Szorzat	16	0,805	0,303	0,353
Harmonikus közép	18	0,777	0,350	0,407
Dice	15	0,712	0,288	0,365
Jaccard	17	0,720	0,329	0,373

legjobb tíz kulcsszóra támaszkodni, szemben azzal a megközelítéssel, amikor a dokumentum összes n-gramjai közül a tíz legmagasabb tf-idf értékűvel lettek jellemezve az egyes dokumentumok. A kulcsszó alapú megközelítés javára írható az is, hogy annak használata mellett a kialakuló közösségek száma közelebbi volt az MSzNy korpuszban beazonosított 27 önálló téma mennyiségéhez.

Mérési eredményeink alapján a dokumentumpárok hasonlóságának súlyozására az átfedésben álló kulcsszavak jóságmértékének harmonikus közepet használó eljárás mondható a legjobbnak (mind az egyszerű n-gramokon, mind pedig a kulcsszavakon alapuló módszer esetében). Ez egyébként megegyezik előzetes várakozásainkkal, hiszen más megközelítések vagy egyáltalán nem hasznosítják a kulcsszavak jóságának mértékét (pl. Dice-együttható), vagy valamilyen értelemben túl szigorúnak (pl. Min), esetleg túl megengedőnek (pl. Max) mondhatók.

További előnyként mutatkozik, hogy a szótár mérete – vagyis azon kifejezések száma, amelyek a dokumentumok közötti összeköttetésekért felelhetnek azzal, hogy legalább egy dokumentumban szerepelnek – a kulcsszavakat figyelembe vevő módszer esetében 2079, míg a dokumentumokban szereplő n-gramokat alapul vevő eljárás esetében ennek több, mint 65-szöröse, 133754 volt.

Ez utóbbi érték természetesen nem azon kifejezések száma, amelyek mind felelősek lehettek két dokumentum közötti hasonlóság megállapítására az n-gram alapú modellben, hiszen dokumentumonként legfeljebb tíz kifejezés lehetett csupán ilyen, a korpusz általunk vizsgált részét alkotó dokumentumok száma pedig kevesebb, mint 400 volt. Ugyanakkor ahhoz, hogy a dokumentumonkénti legjobb tíz tf-idf értékű kifejezés meghatározható legyen, ismernünk kellett az összes, a korpuszban leírt kifejezéssel kapcsolatos előfordulási statisztikát. Ezzel szemben a kulcsszavak meghatározása aktuálisan mindig csak egy dokumentum alapján történt esetünkben, vagyis a szótárt képző kifejezések meghatározása dokumentumonként, egymástól függetlenül történhetett, így minden dokumentum esetében elegendő volt csupán az azt leginkább jellemző tíz kulcsszót tárolni.

7. Konklúzió és további munka

Jelen munkában az MSzNy cikkarchívumának automatikus kulcsszavazását és a kulcsszavazáson alapuló klaszterezését vizsgáltuk. A dokumentumokból épített

hasonlósági gráf particionálására, és így a témájukban koherens diszjunkt részkorpuszok detektálására alkalmasabbnak bizonyult az a megközelítés, amely az automatikusan meghatározott kulcsszavakkal jellemzi az egyes dokumentumokat, mint az n-gram alapú modell. A közös kulcsszóval rendelkező dokumentumok hasonlóságának számszerűsítésére pedig az átfedő kulcskifejezések kulcsszavazó modell által predikált valószínűségeinek felhasználása mutatkozott célravezetőnek (szemben pl. az egyszerű tf-idf mutató használatával).

Munkánk során elkészült a korpusz klaszterezésének egy interaktív online vizualizációja is, amely elérhető a `rgai.inf.u-szeged.hu/DocViewer` oldalon.

A dokumentumok kulcsszavai, illetve a klaszterek hasznos segítséget nyújthatnak számos további (pl. információ-visszakereső) alkalmazás számára, valamint az egyes részkorpuszok (közösségek) méretének változásának időbeli dinamikájának vizsgálatán keresztül lehetőség nyílik a különböző részterületek fontosságának alakulásának monitorozására, trendkövetésre, melyeket a jövőbeli kutatásaink során mélyebben tervezünk vizsgálni.

Köszönetnyilvánítás

Jelen kutatást a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt támogatta az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett.

Hivatkozások

1. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: Kea: Practical automatic keyphrase extraction. In: ACM DL. (1999) 254–255
2. Nguyen, T.D., Kan, M.Y.: Keyphrase extraction in scientific publications. In: Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers. ICADL'07, Berlin, Heidelberg, Springer-Verlag (2007) 317–326
3. Turney, P.: Coherent keyphrase extraction via web mining. In: Proceedings of IJCAI '03. (2003) 434–439
4. Berend, G.: Opinion expression mining by exploiting keyphrase extraction. In: Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, Asian Federation of Natural Language Processing (2011) 1162–1170
5. Farkas, R., Berend, G., Hegedűs, I., Kárpáti, A., Krich, B.: Automatic free-text-tagging of online news archives. In: Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence, Amsterdam, The Netherlands, IOS Press (2010) 529–534
6. Ding, Z., Zhang, Q., Huang, X.: Keyphrase extraction from online news using binary integer programming. In: Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, Asian Federation of Natural Language Processing (2011) 165–173
7. Berend, G., Farkas, R.: Kulcsszókinyerés magyar nyelvű tudományos publikációkból. In: MSzNy 2010 – VIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 47–55

8. Gupta, S., Manning, C.: Analyzing the dynamics of research by extracting key aspects of scientific papers. In: Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, Asian Federation of Natural Language Processing (2011) 1–9
9. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In Tanács, A., Vincze, V., eds.: MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2013) 368–374
10. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69**(2) (2004) 026113+
11. Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: Maximizing Modularity is hard. (2006)
12. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10) (2008) P10008+
13. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 410–420